Archuby, Gustavo; Caprile, Lorena; González, Claudia; Jorquera, Israel; Merlino, Cristian; Pichinini, Mariana; Romero, Roxana

Medición de uso en repositorios digitales : Hacia la construcción de un marco de referencia argentino

III Jornadas de Intercambio y Reflexión acerca de la Investigación en Bibliotecología

28 y 29 de noviembre de 2013

CITA SUGERIDA:
Archuby, G.; Caprile, L.; González, C.; Jorquera, I.; Merlino, C.; Pichinini, M.; Romero, R. (2013) Medición de uso en repositorios digitales : Hacia la construcción de un marco de referencia argentino [en línea]. III Jornadas de Intercambio y Reflexión acerca de la Investigación en Bibliotecología, 28 y 29 de noviembre de 2013, La Plata, Argentina. En Memoria Académica. Disponible en: http://www.memoria.fahce.unlp.edu.ar/trab_eventos/ev.3363/ev.3363.pdf

Actas de las 3ª Jornadas de Intercambios y Reflexiones acerca de la Investigación en Bibliotecología, La Plata, 28-29 de noviembre de 2013. La Plata: Facultad de Humanidades y Ciencias de la Educación de la Universidad Nacional de La Plata, 2013. ISSN 1853-5631

Medición de uso en repositorios digitales. Hacia la construcción de un marco de referencia argentino

Gustavo Archuby[1]; Lorena Caprile[2]; Claudia González[3], Israel Jorquera[2]; Cristian Merlino[4]; Mariana Pichinini[1]; Roxana Romero[5]

[1]Universidad Nacional de La Plata (UNLP). Facultad de Humanidades y Cs. de la Educación (FaHCE), La Plata, Argentina. [2]Universidad Nacional de La Plata (UNLP). Facultad de Cs. Naturales y Museo (FCNyM), La Plata, Argentina. [3]Instituto de Investigaciones en Humanidades y Cs. Sociales - IdIHCS (UNLP-CONICET)., La Plata, Argentina. [4]Universidad Nacional de Mar del Plata (UNMdP). Facultad de Cs. Económicas y Sociales (FCEyS), Mar del Plata, Argentina. [5]Centro de Investigaciones Ópticas - CIOp (CONICET-CIC), La Plata, Argentina. e-mail: cgonzalez@fahce.unlp.edu.ar

Resumen. Este trabajo presenta las actividades realizadas por el denominado Grupo Métricas encargado del desarrollo de uno de los objetivos específicos del proyecto Investigación y Desarrollo en Repositorios Institucionales: aplicaciones y experiencias en universidades de la región bonaerense (PICTO-2010-0149 - 2012/2013). Dicho objetivo contempla el estudio y análisis de métricas de uso de objetos digitales en Repositorios Institucionales para la definición de un marco de referencia que pueda aplicarse en el contexto nacional. Se resumen aquí los principales resultados obtenidos detallando, en primer término, las conclusiones a las que se arribó luego de realizar el relevamiento de iniciativas y proyectos internacionales. Luego, se describe el conjunto de indicadores básicos a ser calculados y su agrupación por grado de complejidad. En tercer término, se explicitan las principales decisiones que se tomaron en torno al desarrollo de un aplicativo de recolección y procesamiento de datos de uso, y se justifica la definición de los requerimientos que se realizó. Finalmente, se concluye con las perspectivas que este tipo de proyectos demandará en el futuro.

1. Introducción

Los repositorios digitales, la denominada ruta verde del acceso abierto, son uno de los canales estratégicos fundamentales para difundir de manera libre y gratuita la creación intelectual generada por una comunidad académica y por ende, para acelerar el avance de la ciencia e impulsar el desarrollo colectivo. Desde los últimos años, las instituciones de educación superior y de ciencia y tecnología comprometidas con la iniciativa de acceso abierto han promovido el desarrollo y la puesta en marcha de repositorios digitales institucionales para disponer su producción académica y científica en abierto, en favor de la propia comunidad científica y de la sociedad toda.

Como canales de diseminación de conocimiento, los repositorios cumplen un rol clave en la política científica de las universidades y centros de investigación y desarrollo. Por

un lado, mejoran la visibilidad y el acceso a los frutos de las actividades de investigación y enseñanza y, por otro, maximizan la repercusión y el impacto potencial de dichos contenidos. Asimismo, no hay que olvidar que los repositorios también cumplen un papel importante como medio de preservación del patrimonio intelectual.

Ahora bien, para demostrar su valor y garantizar su sostenibilidad en el tiempo no basta con defender retóricamente los principios y beneficios del acceso abierto, se requiere presentar evidencia objetiva que refleje el uso de los contenidos digitales dispuestos en abierto. Esa evidencia son los datos de uso que, por lo general, suelen ser presentados de forma agregada en estadísticas de uso. En el contexto de servicios de información web (repositorios y bibliotecas digitales), la noción de uso puede ser interpretada a través de transacciones de acceso, donde cada transacción representa de forma unívoca un evento de tipo solicitud-respuesta entre un usuario -que peticiona un recurso-, y un servicio de información -que contesta el pedido-.

La disponibilidad de estadísticas de uso es, sin duda, un insumo informativo y valioso para la toma de decisiones respecto a múltiples aspectos asociados al funcionamiento, la promoción y la aceptación de un repositorio digital. Su disponibilidad beneficia, en sus diferentes roles, tanto a los gestores de repositorios como a los autores que depositan en ellos sus obras (Bernal y Pemau-Alonso, 2010).

Desde luego, la necesidad de proveer estadísticas de uso consistentes y confiables propone nuevos desafíos a afrontar. Es en esta instancia donde se empiezan a plantear cuestiones metodológicas vinculadas con el registro, la recolección y el procesamiento sistemáticos de datos de uso de objetos digitales.

Estas cuestiones atienden a planteamientos tales como: qué datos se registran y cómo se procesan, y qué métricas se utilizan y cómo se calculan. La granularidad o grado de detalle de las transacciones de acceso que se pueden obtener de los repositorios digitales es tan rica que nos permite trabajar a nivel de ítem individual, lo cual es fundamental para brindar una imagen confiable del uso de cada objeto (artículo, documento de conferencia, tesis) de la colección.

Por supuesto, si además queremos agregar y consolidar datos de uso de múltiples repositorios, es necesario implementar una práctica de trabajo consistente y homogénea, puesto que a falta de estándares internacionales se debe asegurar un práctica común y unificada (Merk, 2008). De esta manera, dispondremos de indicadores uniformes que

expresen y midan las mismas facetas de análisis, requerimiento indispensable para llevar adelante cualquier tipo de evaluación y comparación posterior.

Si bien algunos países ya han instrumentado proyectos nacionales para la recolección y sistematización de estadísticas de uso de repositorios digitales, la situación es, en términos generales, incipiente a nivel mundial. La Argentina presenta, considerando el creciente número de repositorios creados (OpenDOAR, 2013) y la reciente puesta en marcha del Sistema Nacional de Repositorios Digitales (MinCyT, 2013), condiciones favorables para proponer un marco de referencia o iniciativa que trate de abordar esta cuestión en el ámbito nacional.

2. Objetivo

Con el objeto de afrontar este desafío, el Grupo Métricas del proyecto Investigación y Desarrollo en Repositorios Institucionales: aplicaciones y experiencias en universidades de la región bonaerense (PICTO-CIN 2010-0149) se propuso, por un lado, arribar a una práctica estandarizada de recolección y procesamiento de datos de uso de objetos digitales almacenados en repositorios institucionales y, por otro, definir un conjunto de indicadores que reflejen el uso de dichos objetos o sus representaciones. En el trabajo se presentan los resultados alcanzados en pos de consensuar un marco de referencia a nivel nacional.

3. Materiales y método

3.1. Relevamiento de iniciativas. Estado de situación

Varias iniciativas a nivel internacional están trabajando en el desarrollo de estándares para la elaboración de estadísticas de uso, a través de una serie de normas uniformes que pauten las características que deben poseer los informes de estadísticas de uso para la recopilación, intercambio y análisis de los datos de uso de repositorios. El objetivo es reunir estadísticas de uso factibles de ser comparadas, independientemente de la plataforma de publicación, tipo de recurso, el país de origen, el idioma, y el área temática.

Para la elaboración de este trabajo fueron analizadas:

- COUNTER (*Counting Online Usage of Networked Electronic Resources*).

- EMIS (*E-Metrics Instructional System*).
- KE (*Knowledge Exchange*). *Guidelines for the Exchange of Usage Statistics.*
- MESUR (*MEtrics from Scholarly Usage of Resources*).
- OA-Statistik.
- PIRUS2 (*Publisher and Institutional Repository Usage Statistics*).
- Proyecto NISO Z39.7-201X, *Information Services and Use: Metrics & statistics for libraries and information providers - Data Dictionary.*
- Proyecto ISO/TR 20983:2003, I*nformation and documentation - Performance indicators for electronic library services.*
- SURE 2 (*Statistics on the Use of Repositories*).

Algunos de los indicadores más frecuentes que se mencionan en los distintos proyectos son:

- Fecha y hora en el que se produjo el evento siguiendo el formato ISO 8601.
- URL del archivo o del registro de metadatos que se solicita.
- Solicitante: el usuario que ha enviado la solicitud de un archivo o de un registro de metadatos.
- Tipo de solicitud: distingue entre la descarga de un archivo o la visualización de un registro de metadatos.
- Identificador de artículo.
- Número de descargas de textos completos.
- Número de vistas de abstracts y de registros.
- Cantidad de usuarios.
- Procedencia geográfica.
- Ítem más popular.
- Número de solicitudes exitosas por mes.
- Acceso denegado a ítems de contenido por mes.

3.2. Relevamiento de formas posibles de obtención de datos

La bibliografía que trata el tema de la Analítica Web distingue entre el método

cuantitativo y cualitativo para la recolección de datos. El método cualitativo es utilizado principalmente para la mejora de los sitios, siendo sus técnicas principales los test de usabilidad en laboratorio y las encuestas. Dentro del método cuantitativo -que es el que interesa aquí-, se describen cuatro técnicas principales para la obtención de datos: 1) el uso de registros de archivos de transacciones, conocidos como archivos de *logs*; 2) las balizas web, llamada así por traducción de *web beacon* o también llamada *web bug*, técnica que se basa principalmente en la explotación de las *cookies* de los navegadores; 3) el uso de *JavaScript* incrustado en las páginas mediante etiquetas, la técnica preferida por las empresas que ofrecen servicios de Analítica Web; y por último lo que se conoce como *packet sniffing*, que implica la interposición entre el usuario y el servidor web de una capa especial, que puede ser de software o de hardware, que se encarga de recolectar la información. A los fines de este trabajo fueron discutidas especialmente las técnicas 1, 2 y 3 (Bertot et al., 1997; Weischedel y Huizingh, 2006; Waisberg y Kaushik, 2009; Suneetha y Krishnamoorthi, 2009; Dwyer, 2009; Pani et al., 2011; Verma et al., 2011; Hausmann, 2012; Goel y Jha, 2013).

3.2.1. Uso de archivos de registro de transacciones (*logs*)

Muy temprano, en los inicios de la Web, los desarrolladores notaron que no siempre las cosas ocurrían como se esperaba y que por ello era deseable mantener un registro de la totalidad de transacciones que se llevaban a cabo entre los clientes y el servidor web. El objetivo era poder determinar qué archivos producían errores y en qué momento específico lo hacían. Estos registros, que se almacenan en los llamados archivos de *logs*, son utilizados en la actualidad no sólo para detectar problemas, sino también para tener información precisa sobre algunas de las acciones que los usuarios hacen en nuestro sitio.

Cada una de las transacciones queda representada generalmente en una línea de un archivo de texto, registrando datos como fecha, dirección IP, URL que se solicitó, versión del protocolo, entre otros. El formato en que se registran estos datos se encuentra estandarizado, de manera que los archivos de *logs* que generan los diferentes productos de servidores son iguales siempre que compartan la configuración respectiva. Esta normalización responde al llamado *Common Log Format* (CLF) (*The Apache Software Foundation*, 2012) que está conformado por:

- Dirección IP del cliente.
- Un campo de datos no disponible indicado con un guión (-).
- Identificador de Usuario, en el caso de ser usuario identificado, de lo contrario se mostrará un guión.
- Fecha en formato dd/mm/aaaa:hh:mm:ss zona horaria.
- Petición, compuesta por método http utilizado, recurso solicitado y versión del protocolo.
- Código de estado.
- Cantidad de bytes transferidos.
- Una extensión del formato es el *Combined Log Format* (ECLF) que agrega al anterior dos campos:
    - Referer, donde se muestra el servidor del que proviene el cliente (debería ser la página que contiene un enlace al recurso solicitado).
    - Agente, donde se muestra la información de identificación que el navegador del cliente incluye sobre sí mismo.

Por ejemplo, mientras un registro de un archivo de *logs* tiene el siguiente aspecto:

88.14.128.54 - - [24/Mar/2013:06:42:21 -0300] "GET /art_revistas/pr.4485/pr.4485.pdf HTTP/1.1" 200 343459 "http://www.google.es/search?client=safari&rls=en&q=pdf+la+interpretaci%C3%B3n+de+la+naturaleza+diderot&oe=UTF-8&redir_esc=&hl=es&nfpr=&spell=1&sa=X&ei=TchOUd-FEMOFhQeR_IGIDQ&ved=0CBsQvwU" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_3; es-es) AppleWebKit/531.21.11 (KHTML, like Gecko) Version/4.0.4 Safari/531.21.10"

Su interpretación respecto al estándar es:

IP: 88.14.128.54

Campo de datos no disponible: -

Usuario: -

Fecha: [24/Mar/2013:06:42:21 -0300]

Petición: "GET /art_revistas/pr.4485/pr.4485.pdf HTTP/1.1"

Código de estado: 200

Cantidad de bytes transferidos: 343459

Referer: http://www.google.es/search?client=safari&rls=en&q=pdf+la+interpretaci%C3%B3n+de+la+naturaleza+diderot&oe=UTF-8&redir_esc=&hl=es&nfpr=&spell=1&sa=X&ei=TchOUd-FEMOFhQeR_IGIDQ&ved=0CBsQvwU
Agente: Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_3; es-es) AppleWebKit/531.21.11 (KHTML, like Gecko) Version/4.0.4 Safari/531.21.10

Entonces, considerando que el archivo de texto de *logs* respeta por registro una estructura fija de campos, se pueden realizar procesos de parseo o conversión a base de datos con relativa facilidad. Otro problema diferente es el tipo de información que se puede obtener, tema que será abordado con más en detalle en el apartado siguiente.

3.2.2. Uso de balizas web y *cookies*

Esta técnica consiste en colocar en las páginas web imágenes transparentes de 1 x 1 pixel utilizando una etiqueta HTML img src. Estas imágenes suelen estar almacenadas en un servidor diferente del que aloja la página, generalmente el servidor de un tercero que se ocupará de monitorear el tráfico del sitio. De esta manera, cuando llega una solicitud al servidor de la página se dispara a su vez una solicitud al servidor que contiene la imagen, quién se encargará de enviar la imagen al navegador junto con un código que puede leer las *cookies* y capturar los datos de comportamiento de los visitantes anónimos. El uso más común de esta técnica es el seguimiento del comportamiento de los usuarios a través de diferentes sitios web.

3.2.3. Uso de JavaScript mediante etiquetado de páginas

Otra forma posible de obtención de datos es la introducción de código *JavaScript* en las páginas de los sitios, de manera que cuando un visitante solicita un URL a un servidor web, el servidor devuelve la página incluyendo el código *JavaScript* incrustado. Este código se ejecuta mientras se carga la página y captura datos de diferentes características. La variedad de datos que se pueden capturar es enorme, incluyendo los *clicks* y la posición del cursor, el seguimiento de los movimientos del ratón y las pulsaciones sobre el teclado, el tamaño de la ventana de los navegadores y los *plug-ins* instalados. Además, cualquier información que pueda ser capturada por los archivos de registro también se puede capturar con el etiquetado. Esta técnica es la más usada por

las empresas que ofrecen soluciones de analítica web (Piwik, Google Analytics y Open Web Analytics).

3.3. Consideraciones para la elección de la forma de recolección de datos y tratamiento de los mismos

Resulta vital que se seleccione una forma de recolección de datos que permita extraer la mayor cantidad de información pertinente a nuestros fines, y que a su vez lo haga de la manera más segura y cuidada posible. Por ello, se procedió a analizar las ventajas y desventajas de cada una de las técnicas expuestas, arribando a las siguientes conclusiones:

- La implementación de las balizas web es relativamente sencilla y además elimina naturalmente la necesidad del filtrado de robots ya que estos no ejecutan solicitudes de imágenes. Sin embargo, a menos que se exploten las *cookies* del usuario, los datos que se obtienen son muy limitados. Por otra parte, se sabe que depender de la recolección de datos vía *cookies* es muy arriesgado ya que basta con que el usuario las deshabilite en su navegador o que utilice un programa *antispyware* para que se pierda la información de esa visita. Además, si bien es una técnica recomendada para los casos en los que hay que recolectar información de diferentes servidores, no es un factor a considerar en la presente propuesta, que está pensada para que la recolección la realice cada Repositorio Institucional individualmente.
- La implementación de los *JavaScripts* es también relativamente sencilla, basta con agregar unas pocas líneas de código al pie de la página. Si se decide utilizar los servicios de un proveedor de soluciones de Analítica Web, ni siquiera hay que molestarse en escribir el código, ellos lo proveen, así como los reportes *on-line* resultantes. Esta técnica no presenta problemas con las páginas alojadas en caché, ya que el *JavaScript* se ejecuta independientemente de donde esté alojada la página y si el *script* es propio, se tiene absoluto control sobre los datos que se desean recolectar. Sin embargo, si el usuario tiene deshabilitado el *JavaScript* en el navegador por razones de seguridad o privacidad, la información de ese visitante no se recolecta. Lo mismo sucede con los motores

de búsqueda, que por lo general no tienen la opción disponible, aunque esto puede ser una ventaja para determinados usos al no ser necesario el paso de filtrado. Si bien la posibilidad de utilizar los servicios de empresas es muy tentadora, la realidad es que los datos los poseen ellos y los usuarios de sus servicios deben conformarse con exportar los datos con determinados niveles de agregación y profundidad pre-fijados.

- La recolección de datos basada en los registros de transacciones (archivos de *logs*) se puede iniciar inmediatamente ya que los archivos están a disposición siempre que el administrador del servidor no los haya eliminado. Si bien su principal limitación es que se pierde el registro de acceso a las páginas almacenadas en el caché del ISP, se debe considerar que para el objetivo que se persigue con este trabajo lo que más interesa es el registro de las descargas de los objetos digitales del repositorio, que seguramente no se guardan en *caché*. Por ser el único mecanismo que permite obtener información sobre los motores de búsqueda que acceden a un sitio, es muy usado para realizar estudios relacionados con el posicionamiento en buscadores (SEO).

Teniendo en cuenta estas alternativas, la elección recayó sobre la última técnica, ya que se trata de la única opción que es independiente de las elecciones de seguridad y privacidad que establecen los usuarios al navegar. Si bien la información que se obtiene pareciera en un primer momento muy básica, como se verá más adelante permite el cálculo de diversos indicadores en diferentes niveles de agregación. Para su cálculo, los registros de transacciones deben ser tratados de tal manera que queden únicamente los accesos que:

- tengan código de estado de Hypertext Transfer Protocol 200 (OK) o 304 (No modificado).
- el agente no sea un robot.
- el URL corresponda a una ubicación del repositorio donde se almacenan los objetos digitales (ruta de directorio).
- en caso de haber dos accesos al mismo objeto desde una misma IP en menos de 5 segundos, solo se contará como 1 acceso.

3.4. Propuesta de niveles de medición

Tal como se desprende del apartado anterior, la propuesta es presentar una selección de indicadores factibles de ser calculados con los recursos y base tecnológica que disponen los Repositorios Institucionales de nuestra región. Por ser algunos de ellos de más fácil obtención que otros, se ha propuesto agruparlos en dos secciones. La primera sección, identificada como nivel 1 describe aquellos que pueden obtenerse utilizando como fuente de datos los registros de transacciones, y por lo tanto, accesibles en cualquier institución que posea un repositorio disponible en la web. La segunda (nivel 2), incluye aquellos indicadores que requieren además para su cómputo la obtención de información de otras fuentes, y por lo tanto, conllevan un nivel de complejidad mayor. El criterio utilizado para seleccionar y agrupar los indicadores toma en cuenta, entonces, el nivel de desarrollo de los repositorios e intenta asegurar que todos los servicios puedan relevar el uso de los objetos digitales al menos en un nivel mínimo.

Para el cálculo de los indicadores del nivel 2 se requiere la utilización de información referida a los objetos digitales que no está presente en los registros de transacciones. El elemento que se usa en todos los casos para localizar esta información en otras fuentes es el URL del objeto digital (presente en la petición realizada al servidor web y registrada como transacción), que deberá servir como llave para su identificación. Las fuentes alternativas son aquellas que describen al objeto digital mediante metadatos. Estos pueden estar almacenados en una base de datos jerárquica o relacional, y accesibles mediante alguna interfaz web. Otra fuente alterna que garantiza la homogeneidad de presentación de los metadatos es el servidor OAI-PMH (*Open Archives Initiative - Protocol for Metadata Harvesting*) del repositorio, que funciona como proveedor de datos y que por defecto va a exponer los mismos siguiendo el esquema Dublin Core Simple. Aquí se toma esta fuente como base para realizar el cálculo de los indicadores de nivel 2.

4. Resultados

Para obtener resultados válidos y comparables para cada indicador se brindan las características técnicas para su medición. Los indicadores deben ser calculados de acuerdo a los parámetros delineados para prevenir problemas en la medición que anulen

los datos obtenidos. Para realizar un análisis y comparación integral de la evolución de la actividad de un Repositorio Institucional se deberían medir todos los indicadores, ya sean los correspondientes al nivel 1 o 2.

Los indicadores calculan valores absolutos y porcentajes. En el caso de los indicadores que calculen valores absolutos se deben listar las 20 frecuencias más altas, con sus respectivos valores, con la posibilidad de acceder a la lista completa en una segunda instancia. Cuando el indicador requiera el cálculo de porcentajes se deben mostrar todos los datos intervinientes. Los períodos de tiempo para presentar los resultados de los indicadores son mensuales y anuales.

La estandarización en el cálculo de los indicadores asegura la homogeneidad de los datos obtenidos para observar el desempeño a través del tiempo, para la comparación entre períodos de tiempo y entre Repositorios Institucionales.

A continuación se listan los indicadores establecidos. Por cuestiones de limitación en la extensión del trabajo, no se incluye el detalle de su cálculo, la forma de medición, los niveles de agregación/desagregación y los ejemplos de aplicación de cada indicador, que forman parte del informe final del proyecto de Investigación involucrado.

4.1 Indicadores de primer nivel

4.1.1. Descargas de objetos digitales por procedencia

4.1.1.1. Número total de Descargas (NTD)

Determinar el total de descargas que se han realizado de objetos digitales del repositorio en un período de tiempo.

4.1.1.2. Número de Descargas Internas (NDI)

Determinar la cantidad de descargas de objetos digitales del repositorio que se han realizado desde cualquier página del mismo repositorio, en un período de tiempo.

4.1.1.3. Número de Descargas desde Motores de Búsqueda (NDMB)

Determinar el número de descargas de objetos digitales del repositorio que se han realizado desde la página de resultados de un buscador web, ya sea este generalista u orientado, en un período de tiempo.

4.1.1.4. Número de Descargas Directas (NDD)

Determinar, en un período de tiempo, el número de descargas de objetos digitales del

repositorio que se han realizado de manera directa, sin mediar para el acceso al archivo, motores de búsqueda, sitios web o similares.

4.1.1.5. Número de Descargas desde Otros Servicios (NDOS)

Determinar, en un período de tiempo, el número de descargas de objetos digitales del repositorio que se han realizado desde servicios externos, es decir, aquellas que realiza un usuario que accede desde un sitio que no sea el propio repositorio, un motor de búsqueda o escribiendo el URL en el navegador.

4.1.1.6. Número de Descargas Externas (NDE)

Determinar, en un período de tiempo, el número de descargas de objetos digitales del repositorio que se han realizado desde fuera del mismo. Es decir, incluye las descargas directas (DD), las descargas realizadas desde la página de resultados de un motor de búsqueda (DMB) y las descargas realizadas desde otros servicios (DOS).

4.1.2. Descargas de objetos digitales por zona geográfica

Los indicadores de este punto, que discriminan la descarga por zona geográfica, y los que se describen más abajo relacionados a la visualización de registros bibliográficos por zona geográfica, tienen una limitación importante que está dada por la forma de obtener el país desde el número IP del cliente, ya que los IP que hoy pertenecen a un país, mañana pueden pertenecer a otro, por tanto debería registrarse el país en el momento en que se genera el *log* y no al momento de obtener el indicador.

Por otro lado, a pesar de tener una lista constantemente actualizada con los rangos de IP aún es posible que quede asentado en el registro accesos (*logs*) un IP que no se ha contemplado en la lista de rangos usada en ese momento. Sin embargo es posible identificarlos correctamente con posterioridad.

4.1.2.1. Porcentaje de descargas de objetos digitales por país (PDP)

Determinar, en un período de tiempo, la proporción de descargas de objetos digitales del repositorio que se han realizado desde diferentes países.

4.1.2.2. Porcentaje de descargas nacionales de objetos digitales (PDN)

Determinar la proporción de descargas de objetos digitales del repositorio que se realizan desde dominios pertenecientes al país del repositorio.

4.1.2.3. Porcentaje de descargas extranjeras de objetos digitales (PDE)

Determinar la proporción de descargas de objetos digitales del repositorio que se realizan desde dominios pertenecientes a cualquier otro país que no sea el país de origen

del repositorio.

4.1.3. Visualización de registros bibliográficos

4.1.3.1 Número total de visualizaciones de registros bibliográficos (NTVRB)

Determinar, en un período de tiempo, el número de visualizaciones de registros bibliográficos que se han realizado en nuestro repositorio. Este indicador debería ponerse en relación con el número total de descargas (NTD) para determinar una tasa de uso exclusivamente referencial del repositorio.

4.1.3.2. Porcentaje de visualizaciones de registros bibliográficos por país (PVRB)

Determinar, en un período de tiempo, la proporción de visualizaciones de registros bibliográficos que han realizado usuarios de distintos países.

4.1.3.3. Porcentaje de visualizaciones de registros bibliográficos, nacionales (PVRBN)

Determinar la proporción de visualizaciones de registros bibliográficos que se realizan desde dominios pertenecientes al país del repositorio.

4.1.3.4. Porcentaje de visualizaciones de registros bibliográficos extranjeras (PVRBE)

Determinar la proporción de visualizaciones de registros bibliográficos del repositorio que se realizan desde dominios pertenecientes a cualquier otro país que no sea el país de origen del repositorio.

4.2. Indicadores de nivel 2

4.2.1. Número de objetos digitales no descargados (NODND)

Determinar el uso relativo de los objetos digitales del repositorio en un período determinado.

4.2.2. Número de descargas de objetos digitales por tipo (NDODT)

Determinar la cantidad de descargas que corresponden a algunos tipos de objeto digitales específicos, por ejemplo libros, artículos, tesis, etc.

4.2.3. Número de descargas de objetos digitales por tema principal

Determinar la cantidad de descargas que corresponden a los temas principales de los objetos digitales accedidos.

4.2.4. Número de descargas de objetos digitales por idioma

Determinar la cantidad de descargas que corresponden a los idiomas principales de los

objetos digitales accedidos.

4.2.5. Número de descargas de objetos digitales por año de publicación

Determinar la cantidad de descargas que corresponden a un año específico de publicación de los documentos.

4.2.6. Número de descargas de objetos digitales por título de revista

Determinar la cantidad de descargas que corresponden a una revista determinada, en repositorios con colecciones de revistas.

4.2.7. Número de descargas de objetos digitales por autor

Determinar la cantidad de descargas que corresponden a un autor determinado.

4.2.8. Número de descargas de objetos digitales por versión del documento

Determinar la cantidad de descargas que corresponden a una versión determinada (*draft, pre-print, post-print*).

4.2.9. Número de descargas de objetos digitales por filiación institucional (UNMdP, UNLP, etc)

Determinar la cantidad de descargas que corresponden a las filiaciones de los autores de los objetos digitales más accedidos.

5. Conclusión y avance

Los indicadores de uso antes definidos son una propuesta de normalización de "qué medir" y "cómo medirlo", y su implementación permitirá disponer de estadísticas de uso comparables entre repositorios digitales de diferentes instituciones nacionales. Por estar definidos en función de los datos que se pueden obtener del registro de transacciones del servidor web y del protocolo OAI-PMH, son independientes del software utilizado para desarrollar cada repositorio, lo que garantiza su aplicabilidad en cualquier institución.

Si bien el haber llegado a este punto es un avance significativo, sería deseable además, que se pudiera desarrollar una herramienta que, adaptada a distintos entornos de trabajo, garantice la generación de estos indicadores de forma consistente y homogénea. Una de las funciones principales que debería realizar dicha aplicación es el filtrado de los archivos de *logs*, principalmente para eliminar los accesos de robots, las descargas de archivos que no son del repositorio, los accesos que se producen desde la red local, etc. Una vez realizado el filtrado, la herramienta debería ser capaz de generar una base de

datos que almacene sólo los accesos considerados como válidos. Si bien es recomendable que el histórico de los archivos de accesos se conserve completo, a los fines de esta aplicación es deseable que sólo se almacenen en la base de datos los registros resultantes del proceso de filtrado.

Con los accesos almacenados, la aplicación debería proceder a calcular los indicadores que, como se vio anteriormente, son sencillos conteos o proporciones. También debería permitir que el usuario configure algunos elementos, tales como el lugar donde se encuentra alojado el archivo de *logs*, el rango de fechas por el que quiere calcular los indicadores, la ubicación física de los objetos digitales del repositorio, etc. Una vez realizado el cálculo, la aplicación debería almacenar los datos resumen dentro de la misma base de datos. La visualización de los resultados debería poder hacerse en diversos formatos: csv, xml, txt. Al tener los datos resumen guardados en la base de datos, con facilidad se podrían programar aplicaciones que muestren los reportes *on-line*, de manera similar a como lo hacen los servicios comerciales.

Si bien al momento del desarrollo del presente trabajo esta aplicación es una mera expresión de deseo, la labor realizada por el Grupo Métricas del proyecto PICTO-CIN II 2010-0149 "Investigación y desarrollo de Repositorios institucionales de las Universidades Nacionales de la Región bonaerense. Experiencias y aplicaciones" ha buscado contribuir con avances significativos para su concreción.

Bibliografía

- ANSI/NISO Z39.7-2013, Information Services and Use: Metrics & statistics for libraries and information providers - Data Dictionary. 2013 [en línea] [Citado 29 Jul 2013]. Disponible en World Wide Web: http://z39-7.niso.org/
- Bernal, Isabel y Pemau-Alonso, Julio. 2010. Estadísticas para repositorios: sistema métrico de datos en Digital.CSIC. El profesional de la información [en línea], vol. 19, nº 5, p. 534-544. [Citado 31 Jul 2013]. Disponible en World Wide Web: http://digital.csic.es/bitstream/10261/27913/1/Bernal_Pemau_Estad%C3%ADsticas.pdf
- Bertot, Jhon C. et. al. 1997. Web Usage Statistics: Measurement Issues and Analytical Techniques. Government Information Quarterly, vol. 14, nº 4, p.

373-395.

- COUNTER (Counting Online Usage of Networked Electronic Resources). 2012 [en línea] [Citado 1 Ago 2013]. Disponible en World Wide Web: http://www.projectcounter.org/
- Dwyer, Catherine. 2009. Behavioral Targeting: A Case Study of Consumer Tracking on Levis.com. Proceedings of the Fifteenth Americas Conference on Information Systems, San Francisco, California August 6th-9th 2009 [en línea] [Citado 30 Jul 2013]. Disponible en World Wide Web: http://csis.pace.edu/~dwyer/research/AMCISDwyer2009.pdf
- EMIS (E-Metrics Instructional System). 2004 [en línea] [Citado 2 Ago 2013]. Disponible en World Wide Web: http://emis.ii.fsu.edu/
- Goel, Neha y Jha, C.K. 2013. Analyzing Users Behavior from Web Access Logs Using Automated Log Analyzer Tool. International Journal of Computer Applications (IJCA) [en línea], vol. 62, nº 2, p.29-33. [Citado 30 Jul 2013]. Disponible en World Wide Web: http://research.ijcaonline.org/volume62/number2/pxc3884643.pdf
- Haussman, Verena. 2012. Developing a Framework for Web Analytics. University of Koblenz-Landau.Thesis for Master of Science in Information Management [en línea] [Citado 2 Ago 2013]. Disponible en World Wide Web: http://kola.opus.hbz-nrw.de/volltexte/2012/764/pdf/Master_Thesis_Verena_Hausmann.pdf
- KE (Knowledge Exchange). Guidelines for the Exchange of Usage Statistics. 2010 [en línea] [Citado 31 Jul 2013]. Disponible en World Wide Web: http://www.knowledge-exchange.info/
- Merk, Christine y Windisch, Nils. 2008. JISC Usage statistics review: final report. [en línea] [Citado 12 Jul 2013]. Disponible en World Wide Web: http://repository.jisc.ac.uk/250/
- MESUR (MEtrics from Scholarly Usage of Resources). 2006 [en línea] [Citado 2 Ago 2013]. Disponible en World Wide Web: http://mesur.informatics.indiana.edu/
- MinCyT. 2013. Sistema Nacional de Repositorios Digitales. [en línea] [Citado

10 Jul 2013]. Disponible en World Wide Web: http://repositorios.mincyt.gob.ar/

- OA-Statistik. 2009. Specification: Data Format and Exchange for OA Statistics Version 0.5. [en línea] [Citado 2 Ago 2013]. Disponible en World Wide Web: http://www.dini.de/fileadmin/oa-statistik/projektergebnisse/Specification_V5.pdf
- OpenDoar. 2013. Growth of the OpenDOAR Database - Argentina. [en línea] [Citado 3 Ago 2013]. Disponible en World Wide Web: http://www.opendoar.org/onechart.php?cID=11&ctID=&rtID=&clID=&lID=&potID=&rSoftWareName=&search=&groupby=r.rDateAdded&orderby=&charttype=growth&width=600&height=350&caption=Growth%20of%20the%20OpenDOAR%20Database%20-%20Argentina
- Pani, Saroj K. et al. 2011. Web Usage Mining: A Survey on Pattern Extraction from Web Logs. International Journal of Instrumentation, Control & Automation (IJICA) [en línea], vol. 1, nº 1, p.15–23 [Citado 30 Jul 2013]. Disponible en World Wide Web: http://interscience.in/IJICA_Vol1_Iss1/paper_4.pdf
- PIRUS2 (Publisher and Institutional Repository Usage Statistics). 2009 [en línea] [Citado 31 Jul 2013]. Disponible en World Wide Web: http://www.jisc.ac.uk/publications/reports/2009/pirusfinalreport.aspx
- SURE 2 (Statistics on the use of Repositories). 2009. [en línea] [Citado 31 Jul 2013]. Disponible en World Wide Web: http://www.surf.nl/en/projecten/Pages/SURE.aspx
- SURF. [en línea] [Citado 1 Ago 2013]. Disponible en World Wide Web: http://www.surf.nl/en/Pages/default.aspx
- Suneetha, K.R. y Krishnamoorthi, Raghuraman. 2009. Identifying User Behavior by Analyzing Web Server Access Log File. International Journal of Computer Science and Network Security (IJCSNS) [en línea], vol. 9, nº 4, p. 327–332. [Citado 31 Jul 2013]. Disponible en World Wide Web: http://paper.ijcsns.org/07_book/200904/20090444.pdf
- The Apache Software Foundation. 2012. Apache HTTP Server Project [en línea] [Citado 30 Jul 2013]. Disponible en World Wide Web:

http://httpd.apache.org/docs/1.3/logs.html

- Verma, Vikas; Verma, A.K. y Bhatia, S.S., 2011. Comprehensive Analysis of Web Log Files for Mining. International Journal of Computer Science Issues (IJCSI) [en línea], vol. 8, nº 6, p.199-202. [Citado 29 Jul 2013]. Disponible en World Wide Web: http://ijcsi.org/papers/IJCSI-8-6-3-199-202.pdf
- Waisberg, Daniel y Kaushik, Avinash. 2009. Web Analytics 2.0: Empowering Customer Centricity. SEMJ.org [en línea] vol. 2, nº 1. p 1-7. [Citado 31 Jul 2013]. Disponible en World Wide Web: http://online-behavior.com/sites/default/files/web-analytics-i.pdf
- Weischedel, Birgit y Huizingh, Eelko K.R.E. 2006. Website optimization with web metrics: A case study. 8th International Conference on Electronic Commerce [en línea]. Fredericton, Canada: ACM, p. 463-470. [Citado 1 Ago 2013]. Disponible en World Wide Web: http://aaa.volospin.com/BT606B/Website_Optimization_p463-weischedel.pdf